

1. 概述

本周是我投完稿后的第一周，本周我主要完成了以下工作：

(1) 对自己的论文完成过程进行了回溯，反思研究过程中的不足，把握整个投稿过程中值得肯定的地方

(2) 和自己的导师（罗月童老师）就下一步的工作进行了探讨。我的导师更加倾向于投 7 月底的 AI4Vis，因为时间相对更轻松。我个人倾向于争取投 6 月 31 日的 IEEE ACCESS (CPS Data)，因为能投中会对我的未来产生较大作用。尽管知道时间有些紧张，但自己还是有一小点把握去做成这件事。为确定后续究竟该采用何种计划，同时保证一定的可行性，目前我在努力构思下一篇论文的 idea。

(3) 和孟林浩同学讨论她的投稿任务。经过大量阅读文献后，我对她的投稿（也许）有一定的把握。

2. 论文工作体会

通过完成了我的首篇文章，我有了很大的成长，也有一些心得体会。

1) 我首次体会到在截稿的最后一小时，在大家的帮助下强行把论文完成并投出去的过程。没有别人的帮助，我可能无法完成准时投稿。经历过后，我自己成长了不少，知道了如何和别人共同完成工作。偶尔成为被帮助的那个，感觉也挺好。

2) 体会了完整的科研过程，对如何做研究有了一定的理解。

3) 发现了自己做科研的许多误区

4) 对成为一个独立的研究人员有了一定的感悟，同时也能保证和很多人保持一定的联系。

3. 文献阅读

为了提出我下一个工作的蓝图，同时帮助孟林浩完成论文，我广泛地阅读了一些相关文献。经过文献阅读后，我对孟林浩的论文有了一定的想法。我对自己的研究暂时还缺少足够的把握。

本周我粗读了以下文章：

《Dynamic Bike Reposition: A Spatio-Temporal Reinforcement Learning Approach》

本文提出了一种基于时空强化学习的自行车重定位模型来解决动态自行车重定位问题。用聚类降低车站选择问题的复杂性。用强化学习学习每个区域内的自行车重放置策略。用深度神经网络学习一个最优长期值函数（估值）。我们进一步使用时空剪枝技术降低该问题的复杂度（heuristic）。

《Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification》

本文提出了一个多变量统计分析的方法来分析 Secure 2-party Computation (S2C) 框架下的数据交换安全性问题。

《Analyzing the Training Processes of Deep Generative Models》

本文提出了一种可视分析系统用来更好地理解生成网络的可视化进程。之前的神经网络可视化方法，要么过于密集，造成视觉的混杂，同时无法揭露演变趋势。要么过于抽象，无法揭露演变细节。本文提取了大量的时序数据表达训练的动态过程，并用 line-chart 表示出来。用蓝噪声采样降低混杂，同时保留 outlier。最后使用对齐算法进一步分析错误的原因。通用的方法。

《TOWARDS FEDERATED LEARNING AT SCALE: SYSTEM DESIGN》

本文的团队执行了一个联邦学习的时间部署。他们把实际的联邦学习系统分为：protocol（使用参与者的目标数量、最小目标数量百分比、暂停时间等参数来确定什么时候可以训练）、device、server、analytics、secure aggregation、tools and workflow、application。本文使用的联邦学习算法是 FEDERATED AVERAGING（谷歌）。

《TPFlow: Progressive Partition and Multidimensional Pattern Extraction for Large-Scale Spatio-Temporal Data Analysis》

多维时空数据典型的探索方式是多坐标视图，用户通过交互式的方式。然而这种方式是一个冗长的探索过程。为解决该问题，本文提出了一种方法。该方法将时空数据表示成张量的形式，并用算法自动化地将这些数据划分成同质的部分并提出潜在的 pattern 来比较和视觉概述。该算法最后化提出出来的 pattern 能够表达数据的置信度。基于这个算法，我们提出了一种可视分析框架，自上而下逐步地展示不同的细节等级的多维时空数据探索。

本文的可视化技术按照展示和 interactive 来组织。

但这样的划分技术不足以满足要求，因此设计了一套可视分析系统，以得到最佳的划分。可视化允许挖掘一个不完美的算法。可视化可以是给专人使用的系统。

《A visual analytical approach for transfer learning in classification》(B 区)

迁移学习有效的一个重要假设是源域和目标域有很大的相似性，分析相似性会被复杂的转换关系阻碍，也没有直观转换过程。本文提出了一个视觉沟通和交互技术来支持（基于任务的）迁移学习过程。此外，一个视觉辅助的迁移学习方法被提出在文本分类领域。该方法可交互式地选择合适的任务和数据来完成迁移学习任务。本文首先解决采用怎样的 metric of task relevance。它解决的实际上还是数据相似度和把这些数据用在模型上时对模型效果的影响程度。本文设置了相似度度量，数据的分布情况（即认为这些数据可以帮助人们判断迁

移学习采用哪些数据)，让人交互地使用这些可视化设计来获得更高的准确度（即如何利用这些信息是不确定的，但可以让人去尝试使用、探究来得到经验）。本文一共提供了两种分析的视图。首先是源数据域和目标数据域的相似性判断，决定将哪个数据加入迁移学习中，其次是选择的迁移的源域和目标域再选择需要用到哪些数据去训练。交互式等选择训练数据并观察训练结果。

总而言之，这篇论文把问题分析清楚了，再拆分成两个子问题。

《A Utility-aware Visual Approach for Anonymizing Multi-attribute Tabular Data》

为了防止安全问题，要对数据采取一定的安全措施，将导致数据的可用性降低。本文提出一种可视分析方法，展示当隐私保全操作使用时，数据的可用性有了什么样的变化。本文的重点聚焦于可用性的变化上。本文以两种隐私保全模型为例，介绍了如何度量隐私。本文构造了一颗隐私暴露风险树，用于对隐私处理结构进行建模。关于隐私风险树：自顶向下的聚类过程会逐渐暴露数据的隐私信息。同时采用一些安全手段，如添加噪声会降低使用性。本提出了一种可视分析手段，用于分析可用性、安全性在聚类和安全措施采用时的影响。对于达到安全标准的数据，视图能够自动地消减在图上的影响。本文首先确定了 4 个安全性度量指标，然后设计了 PER-Tree 去展示隐私性指标，用可用性表达度量矩阵去展示可用性的变化情况。

《Beyond weber's law: A second look at ranking visualizations of correlation》

散点图在识别数据相关性上显示了无与伦比的性能。还有连续型和离散型散点图。提到了一种识别离散的散点图的方法。

《Differentially Private Distributed Learning for Language Modeling Tasks》

本文解决的核心问题是，如何阻止通用的语言模型在私人数据上 fine-tuning 时忘记“general English”。本文提出的方法还具有通信有效性。本文提出的方法在文本预测的两个指标，perplexity reduction 和 keystroke saving rate，分别获得了 70% 的降低和 8.7% 的提高。为此本文提出了三种模型：1) learning without forgetting (联合概率分布)；2) training with rehearsal (训练集中加入预训练的数据)；3) Server-side model update (在使用 (1) 方法优化的同时，使用 N 个模型的聚合结果作为输出)。

《LEARNING DIFFERENTIALLY PRIVATE RECURRENT LANGUAGE MODELS》

它的想法是，每次我选择用来训练模型的用户都不一样，那么我们不就不知道，每个用户具体的信息了吗？但是随机选择训练，还有另一种破解方法。比如我第 1 轮迭代，选择了 10 个人，第 2 轮迭代，我选择了第 1 轮选择的那 10 个人的同时，额外再加一个人，那么我就可以通过这两轮信息的差异来破解了额外增加的那个人的隐私信息。那么，另一个显然的想法就是每一轮迭代，我都在本轮训练好的模型上面额外加一个噪声。用来保证 differential privacy。我们需要通过某种方式来确定这个噪声的参数，使得这种噪声不会影响到我们模型训练最终的结果。如何确定这个噪声的参数呢？需要确定网络模型梯度的取值边界。然后作者就使用了一系列的 clipping 措施。这样最终保证的结果就是。别人无法通过每一轮。选择用户的不同来推断出用户的敏感信息。

《Federated Multi-Task Learning》—(感觉有细看的价值，重点在那个 node 的 dropping)

本文尝试了联邦学习环境下的多任务学习。本文认为联邦学习特别适合解决统计和系统的挑战。统计挑战：大量 non-IID 数据。但应该有潜在相同的结构。系统挑战：通信情况各式各样（延迟、存储、功耗、中断），需要有不同的差错容忍机制。多任务学习，保证 W 尽可能地不相干。本文引入多任务（把每个 node 看成一个单独的 node）学习模型

$$\min_{\mathbf{W}, \mathbf{U}} \left\{ \sum_{i=1}^I \sum_{j=1}^J \ell(\mathbf{W}_i^T \mathbf{x}_i^j; \mathbf{y}_i^j) + \mathcal{L}(\mathbf{W}, \mathbf{U}) \right\},$$

引入一个相互关系建模矩阵

$$\Omega \in \mathbb{R}^{m \times m}$$

。本文

观察到同时优化 W 和 Ω 不是个 jointly convex 的优化问题。因此采用分步优化，为适应联邦学习环境的系统问题，本文提出了一种独特的 W 更新方法。-----本文接下去没有细看的价值。结果它使用相互关系建模。

其中的收敛分析可以看看，很重要。（拔高的情况下看看）。为解决系统性挑战，本文使用 θ 参数来决定是否丢弃掉该 node。

《Learning Privately from Multiparty Data》。

本文提出了一种是用本地 classifier 生成辅助的未标记的数据来训练网络的方法。

《Generative Knowledge Transfer for Neural Language Models》

本文受到上一篇文章的启发。因为用户的数据不直接用来训练，因此可看做一种隐私保护的
语言适应模型。与其使用用户的数据来训练模型，不如用在用户数据上调整的模型参数来训练一个新的模型（知识迁移方法）。但这种方法仍旧会泄露隐私。于是干脆只使用用户训练好的模型来生成 (generate) 训练数据的标签，这种方式是用 teacher networks 去训练 student networks，这种方式又被称之为 GKT (generative knowledge transfer)，也即是本文的核心贡献，只是用生成数据来训练

4. 时间安排

星期	任务	Duration
周一至周日	调试代码、思考工作汇报内容	9:00 - 12:00 和 14:00 - 23:00，共 12 小时

Work Time: above 50 hours